УДК 004.931

КОНДЕНСАЦІЯ ДАТАСЕТУ ЗОБРАЖЕНЬ ОКТ МЕТОДОМ ВІДПОВІДНИХ ГРАДІЄНТІВ

О. А. Вергелюк[1,a], Н. В. Шаповал[1]

[1] Навчально-науковий інститут прикладного системного аналізу

Анотація

У роботі розглянуто ефективний щодо даних метод машинного навчання – метод конденсації датасету з відповідними градієнтами. Цей алгоритм на стандартних датасетах показав кращі результати, ніж подібні методи. В роботі застосовано метод конденсації датасету з відповідними градієнтами для стандартизованого набору даних зображень тканин ока, отриманих за допомогою оптичної когерентної томографії. Результати моделей, навчених на синтезованих даних, порівняно за точністю із результатами моделей з інших робіт.

Ключові слова: машинне навчання, комп'ютерний зір, згорткові нейронні мережі, класифікація зображень, градієнтний спуск, конденсація даних, оптична когерентна томографія

Вступ

Проблема повної чи часткової втрати зору призводить до значного погіршення якості життя. Найпоширеніша причина полягає в пошкодженні сітківки ока. Але раннє виявлення симптомів та дозволяє призначити препарати, які сприяють відновленню зору. Діагностувати передумови проблем із зором дозволяє оптична когерентна томографія (ОКТ), яку проводять понад 30 мільйонів разів на рік [1].

Значні обсяги даних дозволяють застосовувати методи машинного навчання для виявлення патологій тканин ока. Але із збільшенням кількості даних зростає не тільки точність натренованих моделей, а і обчислювальні витрати на навчання. Особливо ця проблема проявляється тоді, коли виникає потреба періодично оновлювати чи змінювати модель. В таких випадках доцільно використовувати ефективні щодо даних методи машинного навчання. Одним з таких є метод конденсації даних з відповідними градієнтами, який на стандартних датасетах показав кращі результати порівняно з іншими методами [2].

1. Дослідження захворювань ока за допомогою ОКТ

Оптична когерентна томографія ока (ОКТ) – це неінвазивний метод офтальмологічної діагностики, який дозволяє отримати тривимірні зображення тканин ока у високій роздільній здатності [3]. Це одна з найрозповсюдженіших процедур отримання медичних зображень. ОКТ використовує лазерне випромінювання та принцип інтерференції для отримання результатів діагностики. Хоча цей метод було винайдено відносно недавно, у 1991 році, він широко застовується по всьому світу для раннього виявлення передумов сліпоти. Основними причинами втрати зору є діабетичний макулярний набряк та вікова дегенерація жовтої плями.

Із розповсюдженням цукрового діабету поширюються і випадки втрати зору через діабетичний макулярний набряк (ДМН). У наш час ОКТ широко застосовується для раннього виявлення ДМН [4]. Завдяки вчасному виявленню змін сітківки ока в ході діагностики стає можливим призначення anti-VEGF препаратів. Застосування цих ліків на ранніх етапах хвороби відновлює зір.

Вікова дегенерація жовтої плями є найпоширенішою причиною втрати зору літніх людей [5]. Ранніми ознаками захворювання є друзи. Друзи – це відкладення позаклітинного сміття між сітківкою ока та мембраною Бруха. Ці утворення активно вивчаються, але наразі відомо про зв'язок друз із пізніми стадіями дегенерації жовтої плями.

Ще однією причиною розвитку дегенерації жовтої плями є патологія кровоносних судин (хоріоїдальна неоваскуляризація). В ході цього процесу спостерігається надмірне зростання судин в оці. Через це відшаровується пігментований епітелій сітківки, що може призводити до повної втрати зору [6, 7].

Усі ці патології сітківки ока можуть бути виявлені в ході діагностики методом оптичної когерентної томографії. В роботах [1, 8] розгянуто 207130 результатів ОКТ, з яких 108312 відібрано придатними для навчання моделей машинного навчання і ще 1000 обрано для тестування моделей. Цей стандартизований датасет з меншеними розмірами зображень [9, 10] і було обрано для дослідження в ході роботи. На рисунку 1 наведено приклади зображень з набору даних.

[a]oleksandr.verheliuk@mail.com

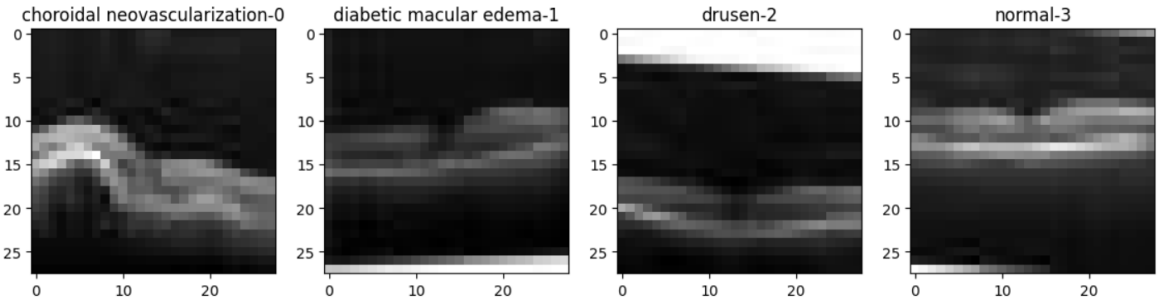


Рис. 1. Приклади зображень з датасету

2. Конденсація даних з відповідними градієнтами

Один з очевидних методів для покращення якості моделі машинного навчання полягає у збільшенні вихідного набору даних для навчання. Але при цьому попередня обробка даних і подальше навчання моделі потребують значних обчислювальних потужностей. І в такому випадку стрімко зростають витрати на навчання моделі: або необхідно облаштовувати необхідну інфраструктуру самостійно, або використовувати сервіси хмарних обчислень. При цьому, в разі необхідності внесення змін у попередні набори даних виникатиме потреба повністю заново перенавчати модель.

Для вирішення цих проблем застосовують ефективні щодо даних методи машинного навчання. Більша частина цих методів пов'язані із вибором з початкової множини репрезентативних в деякому сенсі екземплярів даних, наприклад, метод дистиляції даних [11]. Окрім цього, існують методи, де репрезентативні екземпляри даних генеруються на основі початкового датасету. В даній роботі розглянуто один з таких методів конденсації даних з відповідними градієнтами [2]. За результатами попередніх досліджень цей метод дає кращі результати порівняно із методами, які обирають репрезентативні екземпляри.

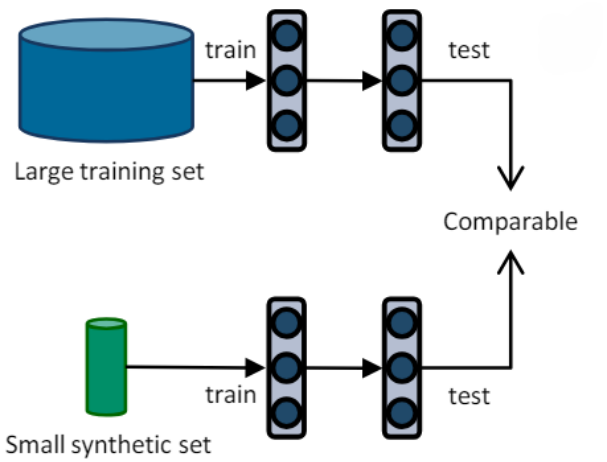
Ідея цього методу полягає в тому, щоб на основі великого початкового датасету згенерувати менший синтетичний датасет, який буде співставний із початковим за результатами натренованих моделей. В ході цього процесу навчаються дві моделі: одна на початковому повному датасеті, а інша на синтетичному наборі. Протягом такого навчання порівнюються градієнти обох моделей і на основі результатів порівняння вдосконалюється синтетичний датасет. Суть та схема методу наведено на рис. 2.

Це має досягатися за рахунок того, що в ході тренування мереж на обох датасетах градієнти параметрів мають мінімально відрізнятися один від одного. Наведемо більш формалізовану постановку задачі.

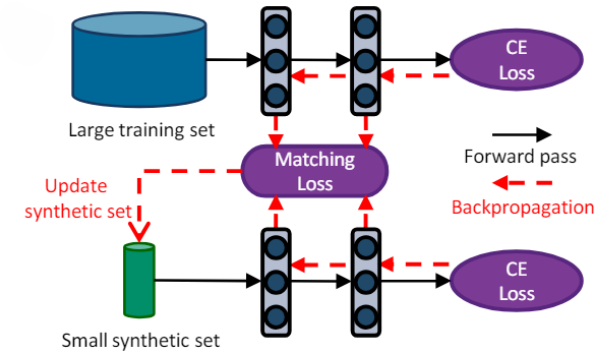
Зазвичай при тренуванні нейронних мереж їх параметри мінімізується з наступної рівності:

$$\theta^{\tau} = \arg\min_{\theta} L^{\tau}(\theta), \qquad (1)$$

де τ – це тренувальний набір даних, θ – це ваги мережі, $L^{\tau}(\theta)$ – це функція втрат (ми використовуємо



(а) Суть конденсації даних



(б) Схема конденсації даних

Рис. 2. Схема методу конденсації даних з відповідними градієнтами

категоріально крос-ентропію). Ми хочемо створити такий синтетичний набір даних S, $|S| \ll |T|$. Для мережі, яка тренуватиметься на синтетичній вибірці:

$$\theta^{S} = \arg\min_{\theta} L^{S}(\theta) \qquad (2)$$

Оскільки синтетичний набір на декілька порядків менший за початковий, то умова (2) оптимізується значно швидше за (1). При цьому результат навчання мереж має бути співставний. У [2] запропоновано таке формулювання умови отримання порівняної продуктивності мереж на повному та синтетичному наборі:

$$\min_{S} D(\theta^{S}, \theta^{T}), \qquad (3)$$

де $D(\theta^S, \theta^T)$ – відстань між наборами ваг. Така оптимізація пристосовує синтетичний датасет до початкових значень ваг моделі. Для пристосування вибірки до можливого розподілу початкових ваг P_{θ_0} застовується формула рівняння (3):

$$\min_S E_{\theta_0 \sim P_{\theta_0}}[D(\theta^S(\theta_0), \theta^T(\theta_0))] \tag{4}$$

Як компроміс між точністю такої оптимізації та швидкістю застосовується спеціальний алгоритм із заданою кількістю ітерацій ζ:

$$\theta^S(S) = \text{opt-alg}_\theta(L^S(\theta), \zeta) \tag{5}$$

Конденсація даних із відповідними градієнтами вимагає не тільки рівність ваг мереж після навчання, а й в ході самого процесу навчання. Тоді умова оптимізації декомпозується і записується в наступному вигляді:

$$\min_S E_{\theta_0 \sim P_{\theta_0}}\left[\sum_{i=0}^{N-1} D(\theta_t^S(\theta_0), \theta_t^T(\theta_0))\right], \tag{6}$$

де N – кількість епох для тренування моделей. У випадку оновлення ваг моделей градієнтним спуском для одного кроку opt-alg правила оновлення будуть в такому вигляді:

$$\theta_{t+1}^S \leftarrow \theta_t^S - \eta_\theta \nabla_\theta L^S(\theta_t^S) \tag{7}$$

$$\theta_{t+1}^T \leftarrow \theta_t^T - \eta_\theta \nabla_\theta L^T(\theta_t^T), \tag{8}$$

Помічено, що $D(\theta^S(\theta_0), \theta^T(\theta_0)) \approx 0$), тому можна замінити θ_t^T на θ_t^S (надалі позначатимемо θ^S як θ), умову оптимізації (6) записати через градієнти:

$$\min_S E_{\theta_0 \sim P_{\theta_0}}\left[\sum_{i=0}^{N-1} D(\nabla_\theta L^S(\theta_t), \nabla_\theta L^T(\theta_t))\right] \tag{9}$$

Власне алгоритм конденсації даних із відповідними градієнтами полягає в пошуку такого синтетичного датасету S, для якого виконується умова (9).

3. Експеримент

Алгоритм конденсації даних із відповідними градієнтами було застосовано для датасету [10]. Цей набір даних представляє широкі можливості для застосування різних медичних вибірок в кількох розмірах. Було протестовано конденсацію даних на чорно-білих зображеннях оптичної когерентної томографії розміром 28×28 пікселів.

В якості моделі обрано згорткову мережу із 3 шарами. Кожен шар має 128 згорткових фільтрів із розміром ядра 3×3, шаром нормалізаії, AvgPooling та функцією активації ReLU. Вихід моделі генерує лінійний класифікатор. Початково синтетичний датасет генерувався із нормально розподіленими випадковими значеннями. Розглянуто роботу алгоритму для створення 10 і 50 синтетичних зображень на кожен з 4 класів з тренувального датасету.

Процес створення синтетичних зображень наведено на рисунку 3. Зображення із синтетичного датасету наведено на рисунку 4. Навчені на датасетах

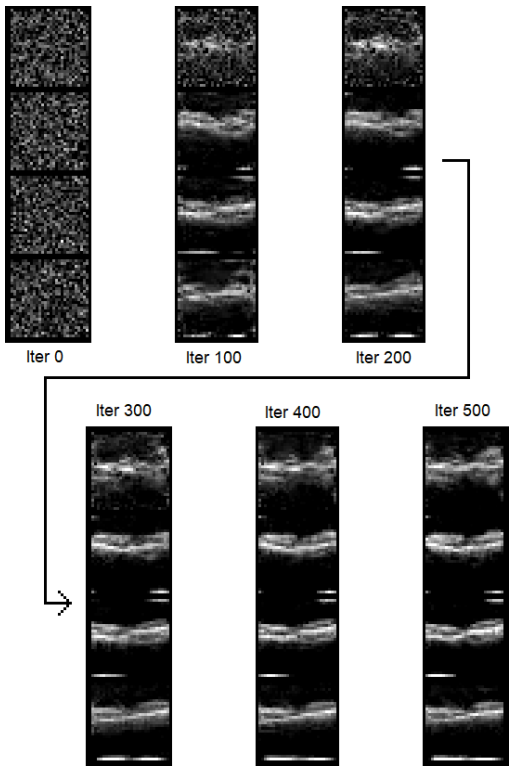


Рис. 3. Процес синтезу зображень кожного класу

із 10 та 50 зображень на кожен клас моделі згорткових нейронних мереж давали показник точності (accuracy) 51.3% та 52.2% відповідно. Це трохи менше, ніж отримано авторами стандартизованого датасету (77.6%), але вони використовували набагато складнішу модель ResNet і найкращі результати вдалося досягти при тренуванні на зображеннях 224×224 пікселів. В експерименті із повнорозмірними зображеннями ОКТ було досягнуто точності 88% за допомогою передавального навчання на мережі ImageNet. Повні результати моделей навчених на синтетичних даних із мережами з інших робіт наведено у таблиці 1.

Таблиця 1. Порівняння результатів тренування на синтетичних даних та всьому датасеті

Model	ipc	Ratio	Accuracy
ConvNet (Our)	10	0.0004	0.513
	50	0.002	0.522
ResNet	–	1	0.763
ImageNet	–	1	0.88

Висновки

Метод конденсації даних із відповідними градієнтами дає середні результати на датасеті із зображеннями оптичної когерентної томографії. Синтетичний набір зображень дозволяє натренувати модель із значно меншими витратами обчислювальних потужностей. Це має особливу цінність, якщо передбачається ситуація, при якій виникатиме потреба змі-

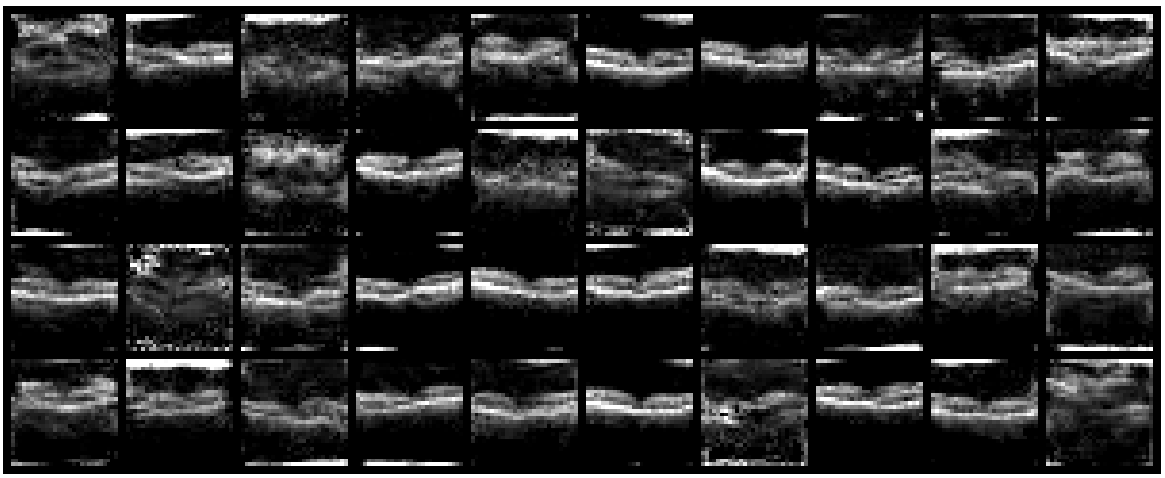


Рис. 4. Синтетичні зображення отримані після конденсації даних (по 10 на кожен клас)

нювати модель відповідно до нових вимог. Щоправда, отримано трохи гірший результат.

На згенерованому синтетичному датасеті із 10 та 50 зображеннями на кожен клас вдалося досягти точності 51.3% та 52.2% відповідно. Це трохи менший результат порівняно з іншими дослідженнями. Покращити результати на синтезованих даних можна шляхом застосування кращої моделі як на етапі створення нової вибірки, так і на етапі тренування кінцевої моделі.

В подальшому можна дослідити ефективність інших методів конденсації даних для зображень ОКТ. При хороших результатах можливе поєдання різних методів в ансамбль для досягнення кращого та збалансованішого результату. Також успішне застовування на стандартизованому датасеті дає підстави використати наведений метод на повнорозмірних зображеннях ОКТ та оцінити ефективність прикладного його застовування для безперервного навчання.

Перелік використаних джерел

1. Identifying medical diagnoses and treatable diseases by image-based deep learning / D. S. Kermany [et al.] // Cell. — 2018. — Feb. — Vol. 172, no. 5. — 1122–1131.e9.
2. *Zhao B.*, *Reddy K. M.*, *Bilen H.* Dataset Condensation with Gradient Matching // International Conference on Learning Representations. — 2021. — URL: https://openreview.net/forum?id=mSAKhLYLSsl.
3. *Varghese M.*, *Varghese S.*, *Preethi S.* Revolutionizing medical imaging: a comprehensive review of optical coherence tomography (OCT) // Journal of Optics. — 2024. — Mar. — DOI: 10.1007/s12596-024-01765-6.
4. Diabetic macular edema: Current understanding, molecular mechanisms and therapeutic implications / J. Zhang, J. Zhang, C. Zhang, J. Zhang, L. Gu, D. Luo, Q. Qiu // Cells. — 2022. — Oct. — Vol. 11, no. 21.
5. Age-related macular degeneration eyes presenting with cuticular drusen and reticular pseudodrusen / J. M. Yoon, D. H. Shin, M. Kong, D.-I. Ham // Sci. Rep. — 2022. — Apr. — Vol. 12, no. 1. — P. 5681.
6. *Yeo N. J. Y.*, *Chan E. J. J.*, *Cheung C.* Choroidal neovascularization: Mechanisms of endothelial dysfunction // Front. Pharmacol. — 2019. — Nov. — Vol. 10. — P. 1363.
7. Identification and clinical role of choroidal neovascularization characteristics based on optical coherence tomography angiography / F. Sulzbacher, A. Pollreisz, A. Kaider, S. Kickinger, S. Sacu, U. Schmidt-Erfurth, Vienna Eye Study Center // Acta Ophthalmol. — 2017. — June. — Vol. 95, no. 4. — P. 414–420.
8. *Kermany D.* Large dataset of labeled optical coherence tomography (OCT) and Chest X-Ray images. — 2018.
9. *Yang J.*, *Shi R.*, *Ni B.* MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis // IEEE 18th International Symposium on Biomedical Imaging (ISBI). — 2021. — P. 191–195.
10. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification / J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni // Scientific Data. — 2023. — Vol. 10, no. 1. — P. 41.
11. Dataset Distillation / T. Wang, J.-Y. Zhu, A. Torralba, A. A. Efros // International Conference on Learning Representations. — 2019. — URL: https://openreview.net/forum?id=Sy4lojC9tm.